Eur J Epidemiol (2010) 25:223–224
DOI 10.1007/s10654-010-9437-5

COMMENTARY

Curbing type I and type II errors

Kenneth J. Rothman

Received: 3 March 2010 / Accepted: 3 March 2010 / Published online: 16 March 2010

The statistical education of scientists emphasizes a flawed approach to data analysis that should have been discarded long ago. This defective method is statistical significance testing. It degrades quantitative findings into a qualitative decision about the data. Its underlying statistic, the *P*-value, conflates two important but distinct aspects of the data, effect size and precision [1]. It has produced countless misinterpretations of data that are often amusing for their folly, but also hair-raising in view of the serious consequences.

Significance testing maintains its hold through brilliant marketing tactics—the appeal of having a "significant" result is nearly irresistible—and through a herd mentality. Novices quickly learn that significant findings are the key to publication and promotion, and that statistical significance is the mantra of many senior scientists who will judge their efforts. Stang et al. [2], in this issue of the journal, liken the grip of statistical significance testing on the biomedical sciences to tyranny, as did Loftus in the social sciences two decades ago [3]. The tyranny depends on collaborators to maintain its stranglehold. Some collude because they do not know better. Others do so because they lack the backbone to swim against the tide.

Students of significance testing are warned about two types of errors, type I and II, also known as alpha and beta errors. A type I error is a false positive, rejecting a null hypothesis that is correct. A type II error is a false negative, a failure to reject a null hypothesis that is false. A large literature, much of it devoted to the topic of multiple comparisons, subgroup analysis, pre-specification of hypotheses, and related topics, are aimed at reducing type I errors [4]. This lopsided emphasis on type I errors comes at the expense of type II errors. The type I error, the false positive, is only possible if the null hypothesis is true. If the null hypothesis is false, a type I error is impossible, but a type II error, the false negative, can occur.

Type I and type II errors are the product of forcing the results of a quantitative analysis into the mold of a decision, which is whether to reject or not to reject the null hypothesis. Reducing interpretations to a dichotomy, however, seriously degrades the information. The consequence is often a misinterpretation of study results, stemming from a failure to separate effect size from precision. Both effect size and precision need to be assessed, but they need to be assessed separately, rather than blended into the *P*-value, which is then degraded into a dichotomous decision about statistical significance.

As an example of what can happen when significance testing is exalted beyond reason, consider the case of the Wall Street Journal investigative reporter who broke the news of a scandal about a medical device maker, Boston Scientific, having supposedly distorted study results [5]. Boston Scientific reported to the FDA that a new device was better than a competing device. They based their conclusion in part on results from a randomized trial in which the significance test showing the superiority of their device had a *P*-value of 0.049, just under the criterion of 0.05 that the FDA used statistical significance. The reporter found, however, that the *P*-value was not significant when calculated using 16 other test procedures that he tried. The *P*-values from those procedures averaged 0.051. According to the news story, that small difference between the reported *P*-value of 0.049 and the journalist's recalculated *P*-value of 0.051 was "the difference between success and failure" [5]. Regardless of what the "correct" *P*-value is for the data in question, it should be obvious that it is absurd to classify the success or failure of this new device

K. J. Rothman (✉)
RTI Health Solutions, Research Triangle Park, NC, USA
e-mail: krothman@rti.org

according to whether or not the *P*-value falls barely on one side or the other of an arbitrary line, especially when the discussion revolves around the third decimal place of the *P*-value. No sensible interpretation of the data from the study should be affected by the news in this newspaper report. Unfortunately, the arbitrary standard imposed by regulatory agencies, which foster that focus on the *P*-value, reduces the prospects for more sensible evaluations.

In their article, Stang et al. [2] not only describe the problems with significance testing, but also allude to the solution, which is to rely on estimation using confidence intervals. Sadly, although the use of confidence intervals is increasing, for many readers and authors they are used only as surrogate tests of statistical significance [6], to note whether the null hypothesis value falls inside the interval or not. This dichotomy is equivalent to the dichotomous interpretation that results from significance testing. When confidence intervals are misused in this way, the entire conclusion can depend on whether the boundary of the interval is located precisely on one side or the other of an artificial criterion point. This is just the kind of mistake that tripped up the Wall Street Journal reporter. Using a confidence interval as a significance test is an opportunity lost.

How should a confidence interval be interpreted? It should be approached in the spirit of a quantitative estimate. A confidence interval allows a measurement of both effect size and precision, the two aspects of study data that are conflated in a *P*-value. A properly interpreted confidence interval allows these two aspects of the results to be inferred separately and quantitatively. The effect size is measured directly by the point estimate, which, if not given explicitly, can be calculated from the two confidence limits. For a difference measure, the point estimate is the arithmetic mean of the two limits, and for a ratio measure, it is the geometric mean. Precision is measured by the narrowness of the confidence interval. Thus, the two limits of a confidence interval convey information on both effect size and precision. The single number that is the *P*-value, even without degrading it into categories of "significant" and "not significant", cannot measure two distinct things. Instead the *P*-value mixes effect size and precision in a way that by itself reveals little about either.

Scientists who wish to avoid type I or type II errors at all costs may have chosen the wrong profession, because making and correcting mistakes are inherent to science. There is a way, however, to minimize both type I and type II errors. All that is needed is simply to abandon significance testing. If one does not impose an artificial and potentially misleading dichotomous interpretation upon the data, one can reduce all type I and type II errors to zero. Instead of significance testing, one can rely on confidence intervals, interpreted quantitatively, not simply as surrogate significance tests. Only then would the analyses be truly quantitative.

Finally, here is a gratuitous bit of advice for testers and estimators alike: both *P*-values and confidence intervals are calculated and all too often interpreted as if the study they came from were free of bias. In reality, every study is biased to some extent. Even those who wisely eschew significance testing should keep in mind that if any study were increased in size, its precision would improve and thus all its confidence intervals would shrink, but as they do, they would eventually converge around incorrect values as a result of bias. The final interpretation should measure effect size and precision separately, while considering bias and even correcting for it [7].

References

1. Lang J, Rothman KJ, Cann CI. That confounded *P*-value (Editorial). Epidemiology. 1998;9:7–8.
2. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol. doi: 10.1007/s10654-010-9440-x.
3. Loftus GR. On the tyranny of hypothesis testing in the social sciences. Contemp Psychol. 1991;36:102–5.
4. Feise RJ. Do multiple outcome measures require *p*-value adjustment? BMC Med Res Methodol. 2002;2:8.
5. Winstein KJ. Boston Scientific stent study flawed. Wall Str J. 2008;August 14:B1.
6. Poole C. Beyond the confidence interval. Am J Public Health. 1987;77:195–9.
7. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer; 2009.